Introduction to Spatial Point Pattern Analysis

by

Stephen L. Rathbun
Department of Health Administration, Biostatistics, and Epidemiology
College of Public Health
University of Georgia
Athens, GA 30605
rathbun@uga.edu

1

References:

Ripley, B.D. (1981). *Spatial Statistics.* Wiley, New York.

Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns,* 2nd Ed. Oxford University Press, London.

Upton G.J.G., and Fingleton, B. (1985). *Spatial Data Analysis by Example.* Wiley, New York.

Waller, L.A., and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data.* Wiley, New York.

Møller, J., and Waagepetersen, R.P. (2004). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman and Hall/CRC, Boca Raton, FL.

2

Definition: A *spatial point pattern* is comprised of the locations of events.

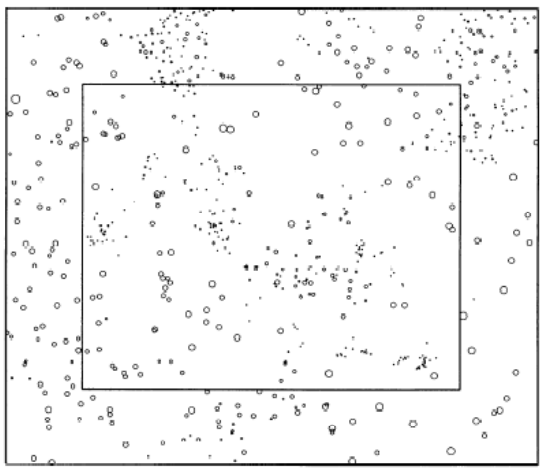
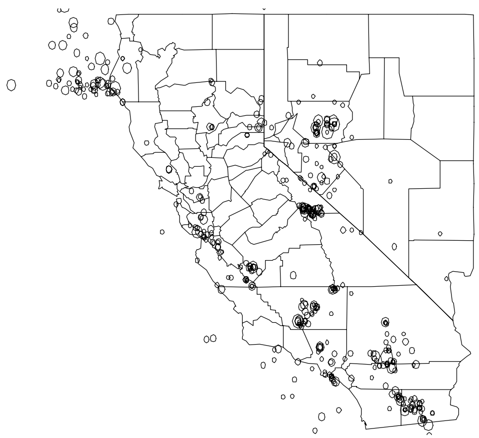


Figure 2. Map of All Longleaf Pines in the 150 × 120 m Study Region B (Inner Rectangle) and the 30 Meter Wide Guard Region $B_+ - B$. The direction north is toward the right side of the page.

3

California Earthquakes



4

Point pattern analysis is primarily concerned with modeling the locations of events, for example the locations of:

- Trees
- Birds' nests
- Ants' nests
- Earthquake epicenters
- Cancer cases
- Galaxies

Objectives: Point Pattern Analysis

1. To determine if the point pattern is completely random;
2. If the pattern is not completely random, fit an explanatory point process model to the data.

Complete Spatial Randomness

Definition: A point pattern is *completely random* if it is realized from a homogeneous Poisson process.

Definition: For a homgeneous Poisson process with intensity λ

1. The number of events (trees) $N(A)$, in a study region A is Poisson distributed with mean $\lambda|A|$

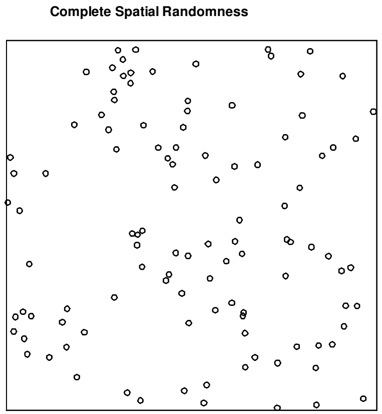
$$\Pr\{N(A) = n\} = \frac{1}{n!}e^{-\lambda|A|}(\lambda|A|)^n$$

2. Conditional on the number of events (trees), the event locations are independently sampled from a uniform distribution on A.

Definition: The *intensity* λ is equal to the mean number of events per unit area.

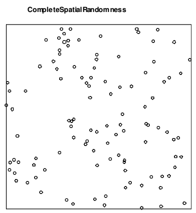
Note: In ecology, the intensity is called the density. In statistics, we use the term intensity to distinguish it from a probability density function.

Completely Random Pattern

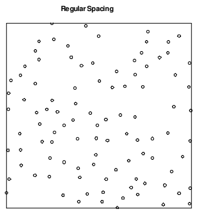


Complete spatial randomness is the null model against which spatial point patterns are often compared.

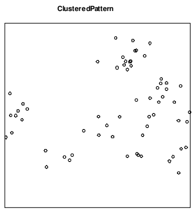
Completely Random



Regular



Clustered



In Ecology:

- Regular spacing may result from intraspecific competition for limited resources;
- Clustered patterns may result from:
  - Clustering of offspring around their parents;
  - Response to a heterogeneous environment.

Ripley's K-Function

Ripley's K-function is the most effective tool for assessing departure from complete spatial randomness.

Definition:

$$K(r) = \frac{\text{Mean number of trees within distance } r \text{ of an arbitrary tree}}{\lambda}$$

Estimation:

$$\widehat{K}(r) = \frac{1}{\widehat{\lambda} N} \sum_{i \neq j} w_{ij} I(d_{ij} \leq r)$$

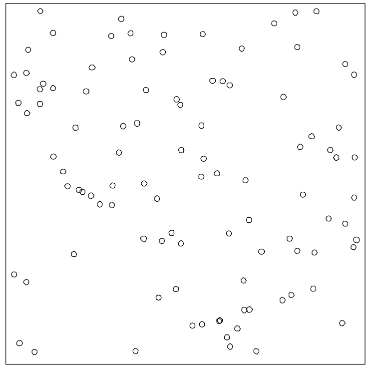
where

$$\widehat{\lambda} = \frac{N}{|A|}$$

is the number of trees in the study region divided by the area of the study region.

What is this?

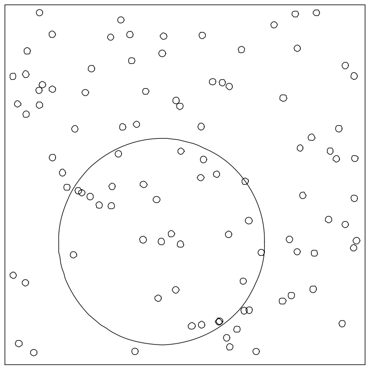
Consider the point pattern of trees:



Here there are 100 trees in a 10×10 region. So

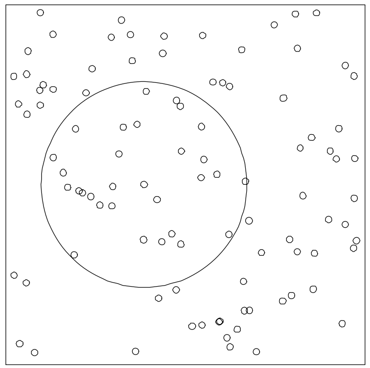
$$\widehat{\lambda} = \frac{100}{10 \times 10} = 1$$

Place a circle of radius r around an arbitrary tree:



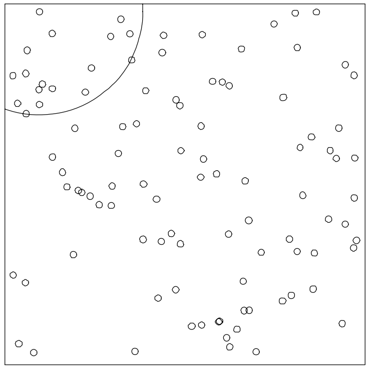
Count the number additional of trees within the circle.

Repeat for each of the remaining trees:



Counting the number of additional trees within each circle.

Edge Correction



For trees close to the edge of the study region, we cannot observe the number of trees within radius r.

Here, we give the neighboring trees a weight w_{ij} equal to one divided by the portion of the circle of radius d_{ij} inside the study region.

13

The results are averaged over all base trees

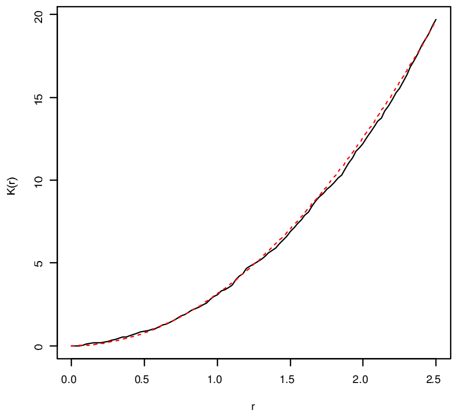
$$\frac{1}{N}\sum_{i\neq j} w_{ij} I(d_{ij} \leq r)$$

and then divided by the estimated intensity $\widehat{\lambda}$ to obtain the estimate

$$\widehat{K}(r) = \frac{1}{\widehat{\lambda} N}\sum_{i\neq j} w_{ij} I(d_{ij} \leq r)$$

14

Plot $\widehat{K}(r)$ against r



Note: Under complete spatial randomness,

$$K(r) = \pi r^2$$

15

Note: Even for strong departures from complete spatial randomness, the difference between the empirical K-function and its expectation under complete spatial randomness is small.

Therefore, a plot of the K-function may not be very informative.

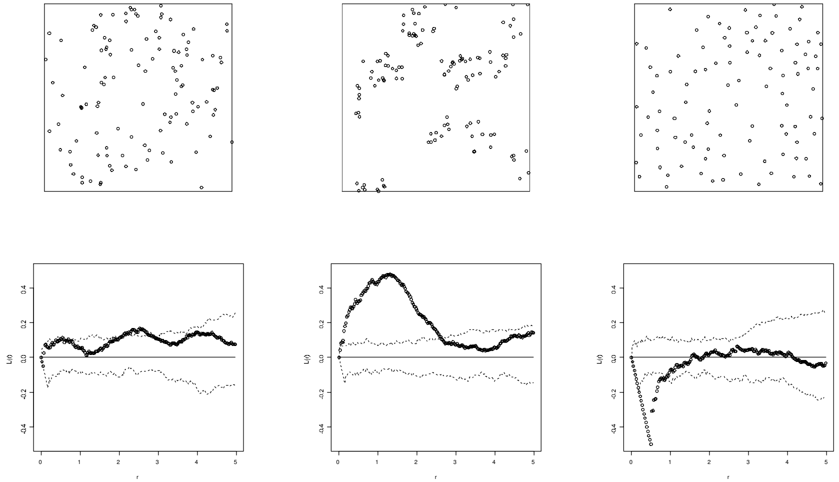
Solution: Linearizing Transformation:

$$L(r) = \sqrt{K(r)/\pi} - r$$

- Under complete spatial randomness
$$L(r) = 0$$
- For clustered patterns
$$L(r) > 0$$
- For regular spacing
$$L(r) < 0$$

16

Completely Random | Clustered | Regular



Note: By plotting the L-function against distance, all scales of pattern can be examined.

17

Point Process Models
Inhomogeneous Poisson Process

Definition: The *intensity* of a point process is

$$\lambda(\mathbf{s}) = \lim_{|d\mathbf{s}|\to 0} \frac{E\{N(d\mathbf{s})\}}{|d\mathbf{s}|}$$

The intensity can be viewed as a local density. Regions with high intensities will tend to contain large numbers of trees, while regions with low intensities will tend to contain few trees.

- $N(d\mathbf{s})$ is the number of trees in a small region $d\mathbf{s}$ surrounding the location $\mathbf{s}$
- $E\{N(d\mathbf{s})\}$ is the mean number of trees in $d\mathbf{s}$
- $|d\mathbf{s}|$ is the area of the region $d\mathbf{s}$
- Thus, the intensity $\lambda(\mathbf{s})$ is the mean number trees per unit area, as a function of location $\mathbf{s}$.

18

Inhomogeneous Poisson Process

The inhomogeneous Poisson process may be used to model the impact of spatial variation in environmental characteristics (e.g., elevation, light intensity, nutrient concentrations) on a point pattern.

Definition: For an inhomgeneous Poisson process with intensity λ

1. The number of events (trees) $N(A)$, in a study region A is Poisson distributed with mean

   $$\Lambda(A) = \int_A \lambda(\mathbf{s})d\mathbf{s}$$

   That is, the probability that the number of events $N(A)$ equal to n is

   $$\Pr\{N(A) = n\} = \frac{1}{n!}e^{-\Lambda(A)}(\Lambda(A))^n$$

2. Conditional on the number of events, the event locations are independently sampled from a probability density function proportional to $\lambda(\mathbf{s})$.

19

Space-Varying Covariates

Let

$$x_1(\mathbf{s}), x_2(\mathbf{s}), \cdots, x_p(\mathbf{s})$$

denote the values of p space-varying covariates at the location $\mathbf{s}$ in the study region A (e.g., elevation, light intensity, nutrient concentrations, etc.).

The impact of these space-varying covariates on a spatial point pattern may be modeled through the intensity function:

$$\lambda(\mathbf{s};\boldsymbol{\beta}) = \exp\{\beta_0 + \beta_1 x_1(\mathbf{s}) + \beta_2 x_2(\mathbf{s}) + \cdots + \beta_p x_p(\mathbf{s})\}.$$

An inhomogeneous Poisson process with the above intensity is called a *modulated Poisson process.*

Reference

Cox, D.R. (1972). The statistical analysis of dependencies in point processes. In P.A.W. Lewis (ed.), *Stochastic Point Processes,* pp. 55-66. New York: Wiley.

20

Parameter Estimation

The *maximum likelihood estimator* is obtained by finding $\widehat{\boldsymbol{\beta}}$ that maximizes the log likelihood:

$$L(\boldsymbol{\beta}) = \boldsymbol{\beta}' \sum_{i=1}^{n} \mathbf{x}(\mathbf{s}_i) - \int_A \exp\{\boldsymbol{\beta}'\mathbf{x}(\mathbf{s})\}$$

where

- $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n$ denote the locations of n trees in the study region A.
- $\mathbf{x}(\mathbf{s})$ = vector of covariates at the location $\mathbf{s}$ in A.

Problem: This requires that the values of the covariates be observed for:

- All of the trees in the study region.
- All locations in the study region.

The former may be impractical, and the later impossible to obtain.

21

Two Approaches:

1. Rathbun (1996) *Biometrics* **52**, 226-242.
2. Rathbun, Shiffman, and Gwaltney (2006) In *Models for Intensive Longitudinal Data.* T.A. Walls and J.L. Schafer (eds.). Oxford.

22

Approach 1

- Sample the covariates at a collection of sites

$$\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$$

- Use kriging to predict the values of the covariates at the locations of the trees, and at the unsampled sites.
- Substitute predicted values into the log likelihood:

$$\widehat{L}(\boldsymbol{\beta}) = n\beta_0 + \beta_1 \sum_{i=1}^{n} \widehat{x}(\mathbf{s}_i)$$

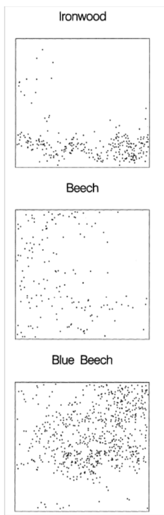
$$- \int_A \exp\{\beta_0 + \beta_1\widehat{x}(\mathbf{s}) + \underbrace{\tfrac{1}{2}\beta_1^2(\sigma^2 - \mathrm{var}(\widehat{x}(\mathbf{s})))}_{\text{Bias Correction}}\}d\mathbf{s}$$

- Find $\widetilde{\boldsymbol{\beta}}$ that maximizes the approximate log likelihood $\widehat{L}(\boldsymbol{\beta})$.

23

Example: Titi Hammock Data

Beech-Magnolia Forest in South Georgia



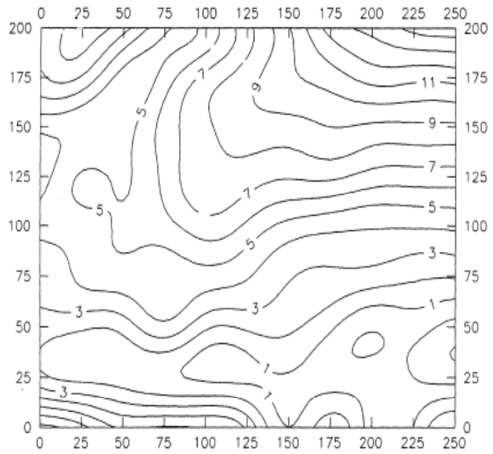


Figure 3. Contour map of predicted elevations in Titi Hammock.

24

Results:

Parameter estimates for a modulated Poisson process with intensity (5.2). Standard errors are given in parentheses.

	No bias correction		Bias corrected	
Species	$\hat{\beta}_0$	$\hat{\beta}_1$	$\tilde{\beta}_0$	$\tilde{\beta}_1$
Bay	−4.3729	−0.9204	−4.4300	−0.8819
	(0.1355)	(0.0936)	(0.1330)	(0.0883)
Beech	−5.4476	−0.0613	−5.4495	−0.0609
	(0.1384)	(0.0262)	(0.1381)	(0.0261)
Blue beech	−5.0711	0.1362	−5.0667	0.1355
	(0.0821)	(0.0113)	(0.0819)	(0.0113)
Holly	−5.2169	0.0161	−5.2164	0.0160
	(0.1099)	(0.0183)	(0.1097)	(0.0182)
Ironwood	−3.2833	−0.7621	−3.3264	−0.7384
	(0.0760)	(0.0432)	(0.0749)	(0.0415)
Magnolia	−5.1454	−0.0277	−5.1462	−0.0276
	(0.1135)	(0.0203)	(0.1132)	(0.0202)
Tulip poplar	−4.8605	−1.0234	−4.9270	−0.9725
	(0.1750)	(0.1356)	(0.1717)	(0.1265)

25

Approach 2

Data Requirements: Covariates are observed

- Locations of the trees

$$\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n$$

- Random locations from the study region

$$\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$$

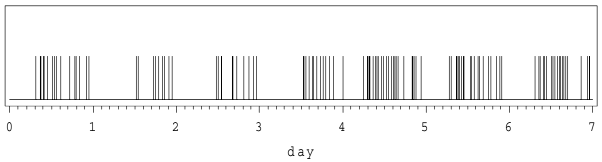
Find $\tilde{\boldsymbol{\beta}}$ that maximizes the approximate log likelihood

$$\widehat{L}(\boldsymbol{\beta}) = \boldsymbol{\beta}' \sum_{i=1}^{n} \mathbf{x}(\mathbf{s}_i) - \frac{|A|}{m} \sum_{j=1}^{m} \exp\{\boldsymbol{\beta}'\mathbf{x}(\mathbf{u}_j)\}$$

26

Example: Ecological Momentary Assessment of Smoking

Times at which cigarettes were lit by a smoker



Time-Varying Covariates

- Negative Affect
- Arousal
- Attention
- Restlessness

27

Results

	Modulate Poisson	
Parameter	Estimate	SE
Intercept	-0.05924	0.00839
Negative Affect	0.01950	0.01077
Arousal	-0.01594	0.01078
Attention	-0.01787	0.01198
Restlessness	0.21017	0.01577

28

Extensions:

- Obtain covariates on a thinned sample of trees. Visit each tree and sample the covariates with known probability p. More generally, p may depend on location.
- Use alternative designs for covariate sample sites:
  - Stratified Random Sample
  - Transect Samples - Random parallel transects, and random sites along each transect.

29